MANTECH Publications
Journal of Artificial Intelligence, Machine Learning and Soft Computing
Volume 9, Issue 2, July-December 2024

Human–Agent Interaction & Trustworthy AI

Sabita S. Bhandari[1], Mohan Pratap Singh[2], Ayesha Kulkarni[3], Devansh Rao[4]
Associate Professor, Professor
Department of Computer Applications
Riverdale College of Engineering
Email ID: *sabitasbhandaribb@gmail.com[1], mohanspratap@yahoo.com[2], ayesha29kulk@rediffmail.com[3]*

Abstract

Human–Agent Interaction (HAI) has become an important field as Artificial Intelligence (AI) systems are increasingly integrated into daily life through chatbots, recommender systems, virtual assistants, and autonomous devices. While AI agents offer efficiency and automation, the issue of trust remains central to user acceptance and ethical deployment. Trustworthy AI involves transparency, fairness, explainability, privacy, accountability, and safety. This paper reviews the relationship between human-agent interaction and trustworthy AI, discussing how user trust is built or broken through interaction design, explainability methods, ethical considerations, and system reliability. It explores models of human trust, design principles for interaction, evaluation techniques, and real-world applications across healthcare, finance, education, and smart systems. The paper also highlights challenges and future directions in creating AI agents that humans can rely upon and collaborate with effectively.

Keywords: *Human-Agent Interaction, Trustworthy AI, Explainable AI, Human-Centered Design, Ethical AI, User Trust, AI Transparency*

INTRODUCTION

Artificial Intelligence agents are now interacting with humans in many forms such as virtual assistants, recommendation engines, automated vehicles, and decision-support tools. These interactions require more than technical efficiency; they require human trust. If users do not trust AI agents, they may reject or misuse them, which limits their potential benefits.

Page 47-58 © MANTECH PUBLICATIONS 2024. All Rights Reserved

Human–Agent Interaction (HAI) studies how humans communicate and collaborate with intelligent systems. Trustworthy AI refers to AI systems that are transparent, fair, reliable, explainable, and safe. When these two areas combine, we obtain systems that are not only intelligent but also socially acceptable.

Trust is a psychological state where a user believes that the agent will act as expected. Lack of trust leads to resistance, while over-trust may lead to blind reliance. Therefore, balanced trust is important in HAI.

2. UNDERSTANDING HUMAN–AGENT INTERACTION (ELABORATED)

Human–Agent Interaction (HAI) is an interdisciplinary area that studies how humans communicate, collaborate, and coexist with intelligent systems. It combines ideas from computer science, psychology, human–computer interaction, cognitive science, and social sciences. The main goal of HAI is to design AI agents that humans can understand, predict, and work with comfortably.

Unlike traditional software, AI agents are adaptive, autonomous, and sometimes unpredictable. Because of this, the interaction between humans and AI is not only technical but also cognitive and emotional. Users do not just "use" AI systems; they **form perceptions**, **build expectations**, and sometimes even **develop social relationships** with them.
HAI includes several important aspects that influence how trust and usability are developed.

2.1 Communication Interfaces (Voice, Text, Gesture)

The way humans communicate with AI agents significantly affects how natural and effective the interaction feels.

- **Voice Interfaces:** Virtual assistants like Alexa or Google Assistant allow conversational interaction. Voice makes interaction feel human-like but may create unrealistic expectations of intelligence.
- **Text Interfaces:** Chatbots and messaging agents rely on text. These systems are common in customer service and education.
- **Gesture and Visual Interfaces:** In robotics and AR/VR systems, gestures, facial expressions, and visual cues are used to communicate with AI agents.

Page 47-58 © MANTECH PUBLICATIONS 2024. All Rights Reserved

A well-designed interface reduces cognitive load and makes the system easier to trust. Poor interfaces may confuse users and reduce confidence in the system.

2.2 Behavioral Responses of AI

AI agents respond to user input based on algorithms and learned patterns. These responses must appear:

- Consistent
- Context-aware
- Timely
- Relevant

If an AI behaves unpredictably, users may feel uncertain and lose trust. For example, a recommendation system that suddenly suggests irrelevant items may appear unreliable.

Behavior also includes **tone**, **language style**, and **response timing**. Friendly, polite, and clear responses tend to increase user comfort.

2.3 User Expectations and Mental Models

Users form mental models about how an AI system works. A mental model is the user's internal understanding of the system's behavior.

For example:

- Users may think a chatbot "understands" emotions.
- Users may assume AI has human-like reasoning abilities.

When AI behavior does not match these expectations, confusion occurs. Proper onboarding, explanations, and feedback help align user mental models with actual AI capabilities.

2.4 Social and Emotional Aspects of Interaction

Humans often treat AI agents as social entities. This is known as **anthropomorphism**—assigning human traits to machines.

Examples include:

- Saying "thank you" to virtual assistants
- Feeling frustrated when chatbots fail
- Trusting AI advice like human advice

Page 47-58 © MANTECH PUBLICATIONS 2024. All Rights Reserved

Emotional design in AI includes:

- Empathetic responses
- Friendly conversational tone
- Personalization

While this improves engagement, it can also create **over-trust** if users believe AI is more capable than it actually is.

Table 1: Components of HAI

Component	**Description**
Interface	Mode of communication (voice, GUI, text)
Perception	How users understand AI behavior
Feedback	How AI responds to user input
Adaptability	Ability to learn user preferences
Transparency	Clarity in AI decisions

3. CONCEPT OF TRUST IN AI SYSTEMS (ELABORATED)

Trust in AI systems is not a single feeling but a combination of beliefs, expectations, and experiences that users develop while interacting with intelligent agents. Unlike traditional software, AI systems make autonomous decisions, learn from data, and sometimes behave in ways that are not fully visible to users. Because of this, trust becomes a central requirement for adoption and effective use.

Trust in AI is often described as **multi-dimensional**, meaning it depends on several factors working together. If even one dimension fails, overall trust may be reduced.

Humans usually build trust gradually through repeated interaction, observation of behavior, and understanding of intentions. In AI systems, however, trust must be **intentionally designed** through system reliability, clear communication, and explainable behavior.

3.1 Competence – "Can the AI do the job correctly?"

Competence refers to the ability of the AI system to perform its intended task accurately and efficiently.

Page 47-58 © MANTECH PUBLICATIONS 2024. All Rights Reserved

Examples:

- A medical diagnosis AI correctly identifying diseases
- A navigation system giving accurate routes
- A fraud detection system identifying suspicious transactions

If AI frequently makes mistakes, users quickly lose confidence. Demonstrating high performance, accuracy metrics, and validated results increases perceived competence.

Design implication: Show performance indicators, confidence scores, and validation results to users when possible.

3.2 Predictability – "Will the AI behave consistently?"

Predictability means that the AI behaves in a consistent and understandable manner over time. Users should be able to anticipate how the system will respond in certain situations.

For example:

- A chatbot providing similar quality responses for similar questions
- A recommendation system not changing preferences randomly

Unpredictable behavior creates uncertainty and reduces trust, even if the system is technically correct.

Design implication: Maintain consistent interaction patterns and avoid sudden unexplained changes in behavior.

3.3 Transparency – "Do users understand how decisions are made?"

Transparency involves providing visibility into how AI decisions are generated. Since many AI models are complex (black-box), users may feel uncomfortable relying on decisions they cannot understand.

Examples:

- Showing why a loan application was rejected
- Explaining why a particular product is recommended
- Providing reasoning behind a diagnosis

Transparency reduces suspicion and increases user confidence.

Design implication: Integrate explainable AI techniques and simple natural language explanations.

Page 47-58 © MANTECH PUBLICATIONS 2024. All Rights Reserved

3.4 Benevolence – "Does the AI act in the user's best interest?"

Benevolence refers to the belief that the AI system will not intentionally harm the user and is designed for user benefit.

For instance:

- Protecting user data privacy
- Avoiding manipulative recommendations
- Providing unbiased suggestions

Users are more likely to trust AI systems that appear supportive rather than exploitative.

Design implication: Communicate privacy policies clearly and avoid deceptive design practices.

3.5 Integrity – "Is the AI ethical and fair?"

Integrity relates to ethical behavior, fairness, and adherence to moral and legal standards.

Examples:

- No bias based on gender or ethnicity
- Compliance with regulations
- Fair decision-making processes

If users suspect bias or unethical behavior, trust deteriorates rapidly.

Design implication: Perform bias audits, fairness testing, and communicate ethical safeguards.

4. PRINCIPLES OF TRUSTWORTHY AI (ELABORATED)

Trustworthy AI is not just about high performance or technical sophistication. It emphasizes the **ethical, transparent, and reliable behavior** of AI systems to ensure that users can trust and safely rely on them. Various international organizations—including the European Commission, IEEE, and OECD—have outlined principles to guide the development of trustworthy AI. These principles aim to align AI design with human values, societal norms, and ethical standards.

4.1 Transparency and Explainability

Transparency ensures that users understand how an AI system works and why it makes specific decisions. Explainable AI (XAI) provides mechanisms to clarify complex models in ways that

Page 47-58 © MANTECH PUBLICATIONS 2024. All Rights Reserved

are understandable to humans.

Key aspects:

- **Algorithmic transparency:** Users know which models and data are used.
- **Decision explainability:** Systems can provide human-readable justifications.
- **Process visibility:** Users can see the steps leading to a decision.

Example: In healthcare, an AI system predicting diabetes risk can explain that high blood sugar levels, age, and BMI contributed to the prediction. This allows doctors and patients to trust and act on the system's output.

Benefits:

- Builds user confidence
- Reduces uncertainty and fear
- Helps detect errors or biases early

Challenges:

- Explaining deep learning models without oversimplifying
- Balancing explanation detail with user comprehension

4.2 Fairness and Bias Reduction

AI systems can inadvertently perpetuate or amplify societal biases if the training data is biased. Ensuring fairness is critical for trust and ethical deployment.

Key considerations:

- **Equitable outcomes:** AI should not discriminate based on race, gender, religion, or socioeconomic status.
- **Bias auditing:** Regular checks for systemic biases in datasets and algorithms.
- **Inclusive design:** Consider diverse populations during model development.

Example: Loan approval AI must ensure applicants from minority groups are evaluated fairly, and not penalized due to historical bias in financial datasets.

Benefits:

- Promotes ethical decision-making
- Enhances acceptance across diverse user groups
- Reduces reputational and legal risks

Page 47-58 © MANTECH PUBLICATIONS 2024. All Rights Reserved

Challenges:

- Defining fairness in complex, multi-objective tasks
- Eliminating historical biases from large datasets

4.3 Privacy and Data Protection

Trustworthy AI requires respecting user privacy and safeguarding personal data. Data misuse can severely undermine user trust.

Key aspects:

- **Data minimization:** Collect only the data necessary for AI functionality.
- **Consent and transparency:** Inform users about data collection and use.
- **Secure storage and processing:** Encrypt and anonymize sensitive data.
- **Compliance:** Follow legal frameworks like GDPR, HIPAA, or local regulations.

Example: A fitness tracking AI must securely store users' health data and allow users to control sharing preferences.

Benefits:

- Builds confidence in AI systems
- Protects individuals from harm
- Supports ethical compliance

Challenges:

- Balancing personalization with minimal data collection
- Ensuring security against cyberattacks

4.4 Accountability

AI systems should have clear lines of accountability. Developers, organizations, and stakeholders must take responsibility for AI decisions, especially in high-stakes applications.

Key aspects:

- **Ownership:** Clear identification of parties responsible for AI outputs.
- **Auditability:** Systems should be traceable and decisions reviewable.
- **Error handling:** Procedures for correcting wrong or harmful AI decisions.

Example: In autonomous driving, manufacturers and software providers are accountable for accidents caused by AI decisions.

Page 47-58 © MANTECH PUBLICATIONS 2024. All Rights Reserved

Benefits:

- Ensures ethical deployment
- Encourages responsible AI design
- Provides recourse in case of failures

Challenges:

- Assigning responsibility in multi-agent AI systems
- Defining legal frameworks for autonomous decisions

4.5 Safety and Reliability

AI systems must function reliably and safely under a wide range of conditions. This principle is particularly crucial for safety-critical applications like healthcare, transportation, and industrial automation.

Key aspects:

- **Robustness:** AI should handle unexpected inputs without failing catastrophically.
- **Redundancy:** Backup mechanisms to prevent harm.
- **Continuous monitoring:** Regular checks for performance degradation.
- **Human-in-the-loop:** Option for human override in risky situations.

Example: An industrial robot AI should stop operations safely if sensors detect human presence in dangerous zones.

Benefits:

- Prevents harm to humans and environment
- Enhances long-term trust
- Supports regulatory compliance

Challenges:

- Designing AI for rare or extreme scenarios

5. Interaction Design for Trust

Interaction design plays a major role in shaping trust.

- Clear instructions and responses
- Consistent behavior
- User control and override options

Page 47-58 © MANTECH PUBLICATIONS 2024. All Rights Reserved

- Error handling and feedback
- Simple explanations in natural language

Agents that communicate politely and clearly are more trusted than opaque systems.

EXPLAINABLE AI IN HUMAN INTERACTION

Explainability bridges the gap between complex algorithms and user understanding.

Explainability Method	Use Case
Feature importance	Loan approval systems
Visual explanations	Medical image analysis
Rule-based output	Expert systems
Conversational explanation	Chatbots

Users feel comfortable when AI can justify its actions.

MEASURING TRUST IN HAI

Trust can be measured using:

- User surveys and questionnaires
- Behavioral analysis (usage patterns)
- Error tolerance levels
- User satisfaction ratings

Psychological scales are often used to assess perceived trustworthiness.

APPLICATIONS OF TRUSTWORTHY HAI

Healthcare

AI assistants support doctors but must explain diagnoses clearly.

Finance

Fraud detection AI must show reasoning for blocked transactions.

Education

Tutoring agents must adapt to student needs transparently.

Autonomous Vehicles

Passengers must trust decision-making of vehicles.

Page 47-58 © MANTECH PUBLICATIONS 2024. All Rights Reserved

CHALLENGES IN HUMAN–AGENT TRUST

- Black-box AI models
- Bias in training data
- Over-reliance on automation
- Misinterpretation of AI outputs
- Ethical dilemmas

Balancing automation and human control is difficult.

ETHICAL CONSIDERATIONS

Ethics play a critical role:

- Avoid manipulation
- Respect user autonomy
- Avoid deceptive design
- Provide informed consent

Ethical guidelines ensure sustainable AI adoption.

FUTURE DIRECTIONS

Future research focuses on:

- Emotion-aware AI agents
- Adaptive trust models
- Regulation and policy frameworks
- Interdisciplinary research combining psychology and AI

CONCLUSION

Human–Agent Interaction and Trustworthy AI are deeply connected. AI systems must be designed with human trust in mind, not only performance. Transparent communication, ethical design, and explainability are key to building systems that humans can rely upon. As AI continues to evolve, maintaining trust will be essential for successful human-AI collaboration.

REFERENCES

1. Siau, K., & Wang, W. (2020). Building trust in artificial intelligence. *MIS Quarterly Executive*.
2. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you? *KDD*.

Page 47-58 © MANTECH PUBLICATIONS 2024. All Rights Reserved

3. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable ML.
4. Floridi, L., et al. (2018). AI4People—Ethical framework for AI.
5. European Commission. (2019). Ethics guidelines for trustworthy AI.
6. Hoffman, R. R., et al. (2018). Metrics for explainable AI.
7. Miller, T. (2019). Explanation in AI: Insights from social sciences.
8. Lee, J. D., & See, K. A. (2004). Trust in automation.
9. Parasuraman, R., & Riley, V. (1997). Humans and automation.
10. Binns, R. (2018). Fairness in machine learning.

Page 47-58 © MANTECH PUBLICATIONS 2024. All Rights Reserved